한국시뮬레이션학회 논문지
Vol. 18, No. 4, pp. 137-148 (2009. 12)

선형회귀모델의 변수선택을 위한 다중목적 유전 알고리즘과 응용

김동일[1] · 박정술[1] · 백준걸[1†] · 김성식[1]

Multi-objective Genetic Algorithm for Variable Selection in Linear Regression Model and Application

Dong-Il Kim · Cheong-Sool Park · Jun-Geol Baek · Sung-Shick Kim

ABSTRACT

The purpose of this study is to implement variable selection algorithm which helps construct a reliable linear regression model. If we use all candidate variables to construct a linear regression model, the significance of the model will be decreased and it will cause 'Curse of Dimensionality'. And if the number of data is less than the number of variables (dimension), we cannot construct the regression model. Due to these problems, we consider the variable selection problem as a combinatorial optimization problem, and apply GA (Genetic Algorithm) to the problem. Typical measures of estimating statistical significance are R^2, F-value of regression model, t-value of regression coefficients, and standard error of estimates. We design GA to solve multi-objective functions, because statistical significance of model is not to be estimated by a single measure. We perform experiments using simulation data, designed to consider various kinds of situations. As a result, it shows better performance than LARS (Least Angle Regression) which is an algorithm to solve variable selection problems. We modify algorithm to solve portfolio selection problem which construct portfolio by selecting stocks. We conclude that the algorithm is able to solve real problems.

Key words : Variable selection, Genetic algorithm, Regression, Multi-objective GA

요 약

본 논문의 목적은 신뢰성 있는 선형회귀모델을 구축하기 위하여 후보독립변수 중 유효변수를 선택하는 알고리즘을 구현하는 것이다. 선형회귀모델을 구축하는데 있어서 데이터 상의 모든 후보독립변수를 포함하는 것은 모델의 통계적 유의성을 감소시킬 수 있으며, 차원의 저주(Curse of dimensionality)를 유발할 수 있고, 데이터의 개수보다 변수의 개수가 많을 경우 모델의 구축이 불가능한 문제점 등이 있다. 이와 같은 문제점을 해결하기 위하여 변수선택의 문제를 조합최적화의 문제로 보고 유전알고리즘(Genetic Algorithm)을 활용하였다. 일반적으로 선형회귀모델의 통계적 유의성을 평가하는 대표적인 통계량으로는 종속변수에 대한 독립변수의 설명력을 나타내는 결정계수(R^2), 회귀식의 통계적 유의성을 검정하는 F통계량, 회귀계수의 통계적 유의성을 검정하는 t통계량, 잔차의 표준오차 등이 있다. 모델의 통계적 유의성은 하나의 통계량으로 표현될 수 없으므로 다양한 기준을 고려한 다중목적식(Multi-objective function)을 가지는 유전 알고리즘을 설계하였다. 설계한 알고리즘의 성능평가를 위하여 다양한 조건을 가정한 시뮬레이션 데이터에 적용하였다. 그 결과 구축한 알고리즘이 유효변수를 판단함에 있어 기존의 대표적인 변수선택 알고리즘인 LARS(Least Angle Regression)에 비해 우수한 성능을 보임을 확인할 수 있었다. 또한, 주가데이터를 이용한 포트폴리오 선택에 적용해 본 결과 우수한 응용문제 해결 능력이 있음을 확인할 수 있었다.

주요어 : 변수선택, 유전 알고리즘, 회귀모델, 다중목적 유전 알고리즘

---

* 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2009-0074458).
2009년 9월 11일 접수, 2009년 11월 2일 채택
[1] 고려대학교 정보경영공학부
주 저 자 : 김동일
교신저자 : 백준걸
E-mail; jungeol@korea.ac.kr

1. 서 론

변수선택이란 입·출력 변수간의 예측모델을 구축하는데 있어 후보로 존재하는 독립변수들 중 어떠한 독립변수를 사용해야 하는 지를 결정하는 것이다. 변수선택은 회귀분석(Regression), 기계학습 (Machine learning), 시계열 예측(Time series prediction), 이진 및 다범주 분류(Two-class or multi-class classification) 등 다양한 예측모델에 이용된다(Guyon 등, 2003).

변수선택의 목적은 예측의 정확성을 향상시키고, 연산의 효율성을 증대시키며, 데이터에 대한 더 나은 이해를 제공하는 것이다. 이러한 목적을 이루기 위하여 변수선택 알고리즘이 고려해야하는 핵심적인 요소는 예측에 유효한 변수들을 유지하면서 예측을 방해하는 유효성이 떨어지는 변수들을 제거하는 것이다.

본 논문은 기본적인 예측모델 중 하나인 선형회귀모델에 적합한 변수를 선택하는데 초점을 맞추었다. 선형회귀모델의 구축에 사용하는 독립변수가 많아지면 종속변수를 예측하는 능력은 향상될 수 있다. 그러나 유의하지 않은 변수가 늘어난다면 회귀모델의 통계적 유의성이 오히려 감소하는 결과를 초래할 수 있다. 통계적 유의성이 낮으면 회귀모델을 모집단 수준으로 일반화시키기 어렵다(김두섭 등, 2000). 또한 독립변수의 수가 데이터의 개수보다 많은 경우는 회귀모델의 구축이 불가능하다. 그러므로 변수선택을 통해 유의하지 않은 독립변수가 회귀모델의 통계적 유의성을 감소시키는 것을 방지하고, 데이터가 부족한 경우에도 모델 구축이 가능하다.

회귀모델에 있어서 통계적 유의성은 종속변수에 대한 설명력을 나타내는 결정계수(R^2), 회귀식의 통계적 유의성을 검정하는 F통계량, 회귀계수의 통계적 유의성을 검정하는 t통계량, 잔차(Residual)의 표준오차 등 여러 가지 기준을 동시에 고려해서 판단해야 한다. 따라서 통계적 유의성이 높은 회귀모델을 구축하는 변수선택을 위해서는 앞서 언급한 기준들을 동시에 고려해야 한다.

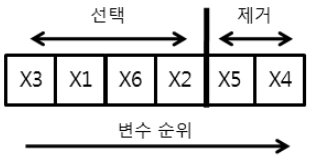


그림 1. LARS와 C_p를 이용한 변수선택의 예

표 1. 조합최적화가 필요한 예

P({A}) = 60	P({A,B}) = 70	
P({B}) = 50	P({A,C}) = 50	P({A,B,C}) = 80
P({C}) = 40	P({B,C}) = 90	

회귀모델을 구축하는데 있어 일반적으로 사용하는 변수선택 알고리즘에는 후진 제거법(Backward elimination), 전진 선택법(Forward selection), 단계적 선택법(Stepwise method) 등의 휴리스틱 알고리즘이 있다(양경숙 등, 2007). 전진 선택법의 하나인 LARS(Least Angle Regression)는 우수한 변수선택 방법으로 알려져 있다(Hesterberg, 2008). 그림 1은 6개의 후보독립변수에 LARS와 일반적으로 사용되고 있는 변수 선택 기준 중 하나인 C_p(Efron 등, 2004)를 적용하여 변수선택을 한 예이다. 그림 1에 적용한 변수선택은 다음과 같은 과정을 따른다. 우선 종속변수에 대한 영향력을 기준으로 후보독립변수들의 순위를 계산하여 정렬한다. 그 다음, 순위가 가장 높은 하나의 독립변수만을 사용하는 변수 조합에서부터 시작하여, 순위가 높은 순서대로 독립변수를 하나씩 추가하며 변수 조합을 만든다. 우선순위가 가장 낮은 변수까지 사용하는 변수 조합을 생성하면 후보독립변수의 개수와 동일한 수의 변수 조합이 생성된다. 그 다음, 생성된 각 변수 조합들의 C_p(식 (17))를 구한다. 그림 1에서는 변수 X4, X5를 제거한 {X1, X2, X3, X6}의 변수조합으로 구축한 모델의 C_p가 가장 낮았기 때문에 최종적인 변수조합으로 결정하였다.

다음은 LARS와 같은 전진 선택법으로 최적의 변수조합을 구할 수 없는 예이다. 종속변수와 함께 주어진 후보 독립변수 A, B, C가 있다고 가정하자. 종속변수에 관한 변수 조합의 영향력을 P(변수 조합)으로 표시한다.

표 1과 같은 상황에 있어서 최적의 변수 조합은 {B,C}이다. 그러나 LARS를 이용한 전진 선택법은 {A}, {A,B}, {A,B,C}의 순서대로 조합할 변수를 찾기 때문에 {B,C} 같은 변수 조합을 찾아내지 못한다. 이렇듯 개별적인 독립변수의 영향력과 조합된 독립변수의 영향력이 다를 수 있기 때문에 정확한 변수선택을 위해서는 조합최적화 문제를 해결하는 알고리즘이 필요하다.

조합최적화 문제로 변수선택 문제에 접근한 알고리즘에는 전역 탐색법(Exhaustive search)이 있다. 전역 탐색법은 특정 기준에 가장 최적화된 변수 조합을 찾을 수 있다는 장점이 있으나 후보독립변수의 개수가 p인 경우 2^p개의 변수 조합에 대해서 판단을 해야 한다. 즉, 후보 변수의 개수가 증가함에 따라 고려해야 하는 경우의 수가 기하급수적으로 늘어난다. Furnival 등(1974)이 제안한 분지탐색(Branch and Bound)도 있지만 이러한 방법도 변수의 수가 어느 정도 이상 넘어가면 제한된 시간 안에

최적해를 찾아내는 것이 불가능해진다. 또한 앞서 언급한 회귀모델의 통계적 유의성에 대한 여러 기준들을 동시에 고려할 수 없다는 단점을 지닌다.

유전 알고리즘은 조합최적화 문제를 해결하는 기법 중 하나이다. 유전 알고리즘은 결정론적 규칙에 따른 탐색이 아닌 유전 연산자와 같은 확률적 변이규칙을 통해 부분 최적 지점을 벗어나 전역 최적 지점을 찾을 수 있는 능력을 가지고 있다(Goldberg, 1989). 또한 고려해야 하는 여러 복잡한 상황을 적합도 함수로 단순화 시킬 수 있다(Mitchell, 1997). 이러한 장점들로 인하여 유전 알고리즘은 knapsack problem, TSP(traveling salesman problem), set covering 등 다양한 조합최적화 문제에 적용되고 있다(Reeves, 1993)(Beasley 등, 1996).

다중목적 유전 알고리즘(Multi-objective genetic algorithm)은 여러 가지 기준을 동시에 고려할 수 있는 최적화 방법이다(Fonseca 등, 1993). 다중목적 유전 알고리즘을 구현하면 회귀모델에서 통계적 유의성을 판정하는데 필요한 여러 가지 기준을 동시에 고려할 수 있어 다양한 평가기준을 고루 만족시키는 모델을 구축하는 변수선택이 가능하다.

유전 알고리즘은 앞서 언급한 변수 선택 휴리스틱 알고리즘들과 비교하여 알고리즘 수행에 오랜 시간이 걸린다. 그러나 본 논문은 변수선택 알고리즘의 수행 속도보다 모델의 통계적 유의성을 높이는 변수선택에 초점을 맞추었다. 왜냐하면 변수선택은 모델을 구축하기 전의 선행과정(Pre-processing)으로 반복적으로 수행하는 과정이 아니고, 일반적으로 변수선택 알고리즘을 수행할 수 있는 충분한 여유시간이 주어지기 때문이다.

이에 따라 본 논문은 회귀모델의 통계적 유의성을 평가하는 기준을 고루 만족시키는 변수 조합을 찾기 위하여 다중목적식(Multi-objective function)을 가지는 유전 알고리즘을 이용하여 MOVS(Multi-Objective Variable Selection)를 구현하였다. MOVS의 성능을 평가하기 위해 두 가지 실험을 수행하였다.

하나는 인위적으로 생성한 데이터를 이용한 LARS와의 비교 시뮬레이션이다. 실제 데이터(Real data)에서는 모델에 직접적으로 영향을 미치는 변수를 알 수 없다. 따라서 유효변수를 판단하는 성능을 분석하기 위해서 유효변수를 알고 있는 시뮬레이션 데이터를 활용하였다.

다른 하나는 주가 데이터를 이용한 포트폴리오 선택에 대한 시뮬레이션으로, 응용문제에 대한 MOVS의 성능을 검증하였다. 포트폴리오 선택은 자산을 투자할 주식과 해당 주식에 투자할 자산의 비율을 결정하는 문제이다. 포트폴리오 선택은 목표 수익과 위험 부담을 동시에 고려해야 하는 특성을 가지고 있다(Luenberger, 1998). 본 논문은 투자 목표 대비 위험부담을 최소로 하기 위한 포트폴리오 선택을 위해 MOVS를 응용하였다. 주가지수에 영향을 미치는 요인이 정확하게 밝혀지지 않아 인위적인 데이터를 생성할 수 없기 때문에 주가 데이터는 과거의 데이터를 사용하는 방식으로 시뮬레이션 하였다. 제안한 알고리즘과의 성능 비교를 위해 포트폴리오 선택에 일반적으로 사용되는 MAD (Mean-Absolute Deviation) 모형과 성능을 비교하였다(Konno 등, 1991).

본 논문의 구성은 다음과 같다. 2장에서는 변수선택의 목적 모델인 선형회귀모델에 관하여 설명한다. 3장에서는 변수선택을 위한 다중목적 유전 알고리즘을 제시한다. 4장에서는 실험을 통하여 제안한 알고리즘의 성능을 검증한다. 5장에서는 결론을 기술한다.

2. 선형회귀모델

선형회귀모델이란 한 개 이상의 독립변수 X와 종속변수 Y에 선형적인 상관관계가 있다고 가정하고 주어진 데이터로부터 관계식을 찾아낸 모델이다. 선형회귀모델은 독립변수와 종속변수와의 관계를 식 (1)과 같이 가정한다.

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{1}$$

이때 ϵ은 평균이 0, 표준편차가 σ인 정규분포를 따르는 독립적인 오차항을 의미한다. 회귀분석을 위해 주어진 데이터를 식 (2)와 같이 독립변수 X의 행렬과, 종속변수 Y의 벡터로 정의하고, 회귀계수를 의미하는 β값들을 벡터로 나타낼 수 있다. p는 독립변수의 수, n은 데이터의 수를 의미한다.

$$X = \begin{bmatrix} x_{11} \cdots x_{1p} \\ \vdots \ \ddots \ \vdots \\ x_{n1} \cdots x_{np} \end{bmatrix},\ Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{2}$$

이를 식 (1)에 적용하면 다음과 같다.

$$Y = X\beta + \epsilon \tag{3}$$

이때 오차를 최소화 하는 회귀계수들을 추정하는 방법은 다음과 같다.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

추정된 회귀계수 $\hat{\beta}$을 가지고 Y값을 추정하는 식은 다음과 같다.

$$\hat{Y}= X\hat{\beta} \tag{5}$$

회귀모델의 통계적 유의성을 검정하려면 분산분석을 수행해야 한다. 추정값 $\hat{Y}$과 관찰값 Y를 이용하여, 분산분석을 수행하기 위해 필요한 총제곱합(*SST* : sum of squares of total), 회귀제곱합(*SSR* : sum of squares of regression), 오차제곱합(*SSE* : sum of squares error)은 다음과 같다.

$$SST= \| Y- \overline{Y} \|^2 \tag{6}$$

$$SSE= \| Y- \hat{Y} \|^2 \tag{7}$$

$$SSR= \| \hat{Y}- \overline{Y} \|^2 \tag{8}$$

선형회귀모델이 실제 모델을 예측하는데 적합한지에 대한 판단은 분산분석을 통해 도출되는 여러 가지 통계량을 고려해서 해야 한다.(김두섭 등, 2000). 대표적인 기준으로 결정계수(R^2), 수정결정계수(Adjusted R^2), 회귀식에 의한 추정값의 표준오차(SEE : standard error of the estimate), 회귀식의 통계적 유의성(F통계량), 각 회귀계수에 대한 통계적 유의성(각 회귀계수의 t통계량) 등이 있다.

결정계수는 종속변수의 총 분산 중 회귀모델로 설명되는 분산의 비율을 나타내는 지수로서 회귀모델의 예측력을 나타낸다. 결정계수는 0과 1 사이의 값을 가지고 1에 가까울수록 회귀모델의 예측력이 높다고 판단할 수 있다. 결정계수는 다음과 같다.

$$R^2= \frac{SSR}{SST} \tag{9}$$

수정결정계수는 표본 자료에서 얻어지는 결정계수가 모집단의 결정계수보다 커지는 경향을 보정하기 위해 자유도를 반영시킨 모결정계수에 대한 추정값이다. 수정결정계수 R^2_{adj}는 다음과 같다.

$$R^2_{adj}= 1- \frac{(n-1)(1-R^2)}{n-p-1} \tag{10}$$

표준오차는 회귀모델로 설명되지 않는 편차의 크기를 의미한다. 표준오차는 다음과 같다.

$$SEE= \sqrt{\frac{SSE}{n-p-1}} \tag{11}$$

추정한 회귀식이 통계적으로 유의한지 검정하기 위한 가설은 다음과 같다.

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ &H_1 : Not\ \ H_0 \end{aligned} \tag{12}$$

식 (12)의 가설을 검정하기 위한 F통계량은 다음과 같다.

$$F_0 = \frac{SSR/p}{SSE/(n-p-1)} \tag{13}$$

또한 독립변수의 회귀계수가 통계적으로 유의한지 검정할 수 있다. i번째 독립변수에 대한 가설은 다음과 같다.

$$\begin{aligned} &H_0 : \beta_i = 0 \\ &H_1 : \beta_i \neq 0 \end{aligned} \tag{14}$$

식 (14)의 가설을 검정하기 위한 t통계량은 식 (15)과 같다. c_{ii}는 $(X^TX)^{-1}$의 i번째 대각원소이다.

$$t_i = \frac{\hat{\beta}_i}{SSE\sqrt{c_{ii}}} \tag{15}$$

회귀모델의 통계적 유의성에 대한 지표들은 앞에서 나열한 식 (10), 식 (11), 식 (13), 식 (15)의 통계량들이며, 통계적으로 유의한 모델은 위의 조건을 고루 만족시켜야 한다.

3. 변수선택을 위한 다중목적 유전 알고리즘

다중목적 최적화 문제는 다수의 목적함수가 상충관계(Trade-off)를 가지는 상황에 있어서 최적해를 찾는 문제이다. 유전 알고리즘으로 다중목적 최적화 문제를 다룬 알고리즘으로는 VEGA(Vector Evaluated Genetic Algorithm)가 있다(Schaffer, 1985). VEGA는 각 목적에 대하여 해집단을 나누어 탐색하는 방법으로, 각 목적에 적합한 해를 동시에 찾아내지만 여러 목적에 동시적으로 적합한 해를 찾지는 못한다(Murata 등, 1995).

Murata 등(1995)은 임의 가중치(Random weight)를 이용해 여러 목적함수를 하나의 적합도 함수로 통일하는 방법을 제안하였다. 그러나 가중합(Weighted sum)을 사용하는 방법은 사용하는 목적함수의 특성에 따라 목적함수의 변환이 필요하다. 또한 목적함수 공간의 인위적인 변환으로, 특정한 해를 찾지 못할 가능성이 생긴다(Deb, 2001).

표 2. MOVS 수행과정

MOVS(*s, q, P(s), P(c), P(m),* E)
s : size of population
q : ranking weight
P(s) : gene selection rate used in initialization
P(c) : crossover rate
P(m) : mutation rate
E : the number used in stoping criterion

G(i, j) : j-th gene of i-th individual
N : non-changed iteration number of best solution
p : number of candidate variable

-initialization … S1
  for(i = 1 to s)
    for(j = 1 to p)
      if(random(0,1) < *P(s)*)
        G(i, j) = 1
      else
        G(i, j) = 0
while(N < E)
-evolution based on each criterion … S2
  random sequence generation : criterion[n] = 0~5
  for(i = 0 to 5)
    -estimation based on each criterion … S3
    criterion[i] = 1 : estimate Adjusted R^2 ranking
    criterion[i] = 2 : estimate F-test p-value ranking
    criterion[i] = 3 : estimate t-test p-value ranking
    criterion[i] = 4 : estimate *SEE* ranking
    criterion[i] = 5 : estimate C_p ranking
    for(c = 0 to p)
      $Fitness(c) = q(1-q)^{Rank(c)-1}$
      $P(select\ c) = \frac{Fitness(c)}{\sum_{i=1}^{p} Fitness(i)}$
    -reproduction … S4
    -crossover and mutation … S5
      for(i = 1 to s/2)
        if(random(0,1) < *P(c)*)
          Two-point crossover
      for(i = 1 to s)
        if(random(0,1) < *P(m)*)
          Simple mutation
-evolution based on total criterion … S6
  estimate ranking based on each 5 criterion
  estimate ranking based on summation of each ranking
  for(c = 0 to p)
    $Fitness(c) = q(1-q)^{Rank(c)-1}$
    $P(select\ c) = \frac{Fitness(c)}{\sum_{i=1}^{p} Fitness(i)}$
  -find best solution of current iteration
  -check stopping criterion … S7
    if(current best solution = prior best solution)
      if(N = E)
        put out current best solution & finish
      else : N = N + 1
    else : N = 0
  -reproduction … S8
  -crossover and mutation … S9
    for(i = 1 to s/2)
      if(random(0,1) < *P(c)*)
        Two-point crossover
    for(i = 1 to s)
      if(random(0,1) < *P(m)*)
        Simple mutation

따라서 본 논문에서는 선형회귀모델의 변수선택을 위한 다중목적 유전 알고리즘인 MOVS를 설계하였다. MOVS는 개별적인 목적함수를 이용한 진화단계와 통합적인 목적함수를 이용한 진화단계가 나누어진 2단계의 진화 과정을 가지는 특징을 가진다. 이때 통합적 목적함수의 구성은 개별적 목적함수에 대한 유전자 개체의 순위를 기반으로 이루어진다. 개별적인 목적함수를 이용한 진화단계에서 각 기준에 대하여 취약한 변수조합이 도태되고, 통합적인 목적함수를 이용한 진화단계에서 모든 기준에 대하여 우수한 변수조합이 선택된다.

MOVS의 수행과정은 표 2와 같다.

다음은 알고리즘의 세부사항이다.

(1) 유전자 개체(해)의 표현 방식 : 변수선택 문제에 있어서 개체의 표현은 전체 후보변수 p의 길이를 가지는 이진수 표현(Binary encoding)이 적합하다(Beasley, 1996). 선택된 변수는 1로 표기하고 선택되지 않은 변수는 0으로 표현한다.

(2) 초기 해집단 생성 (S1) : 초기 해집단의 생성은 변수에 대한 사전 지식이 없다고 가정하고 무작위 초기화(random initialization)법을 이용한다. 이러한 방법은 다양한 개체를 고려할 수 있다는 장점을 지닌다. 이때 해집단의 크기와 각 변수가 초기 해집단 내에서 선택될 확률을 설정해 줄 수 있고 각각 s, P(s)라 한다.

(3) 개별 기준별 진화 (S2) : 적합도 평가에 사용하는 5가지 기준을 독립적으로 사용하여 한 번씩 진화과정을

거친다. 이 과정에서 적합도 평가, 재생산, 변수 판단 및 제거, 교배와 돌연변이 과정이 각 기준에 대해서 한 번씩 수행된다.

사용하는 기준의 순서가 매번 같다면 진화가 한 방향으로만 진행될 수 있기 때문에 사용하는 기준의 순서는 매번 임의로 정해준다. 이 과정을 통해 각 기준을 고루 만족시키는 개체가 남는다. 특히 특정 기준에 대하여 취약한 개체가 있으면 도태된다.

(4) 적합도 평가 (S3) : MOVS는 회귀모델을 평가하는 기준으로 2장에서 언급한 Adjusted R^2, SEE, 회귀식의 F 통계량의 p-value, 각 회귀계수에 대한 t통계량의 p-value의 평균값을 적합도 평가의 기준으로 삼는다.

또한 일반적으로 사용되는 변수선택의 기준 중 하나인 Mallow의 C_p를 적합도 평가 기준으로 활용하였다. Mallow의 C_p는 다음과 같다(Mallows, 2000).

$$C_p = \frac{SSE_k}{SSE/(n-p-1)} - n + 2(k+1) \tag{17}$$

여기서 k는 선택된 변수의 개수, SSE는 모든 후보독립변수로 구축한 모델의 SSE, SSE_k는 선택된 변수 조합으로 구축한 회귀모델의 SSE값이다.

회귀 모델의 유의성에 대한 다양한 평가 기준들은 서로 다른 특성을 지닌다. Adjusted R^2의 값은 클수록 좋은 해이고, 나머지 기준은 낮을수록 좋은 해이다.

이러한 이유로 각 기준을 직접적으로 적합도로 사용하면 각 기준들을 재생산 과정에 공정한 적용이 불가능하다. 따라서 MOVS는 평가 기준에 따른 개체의 순위를 이용한 비선형적 적합도 함수를 이용한다. 해집단 내에서 c번째 해의 순위를 $Rank(c)$라 하고, 가중치를 q라 하면, 적합도는 식 (18)과 같다(문병로, 2003). q는 0과 1 사이의 값을 가진다.

$$Fitness(c) = q(1-q)^{Rank(c)-1} \tag{18}$$

이때 q값이 클수록 순위에 따른 적합도의 차이가 증가한다.

(5) 재생산 (S4, S8) : 재생산은 목적에 적합한 개체일수록 생존 확률을 높여 다음 세대에 목적에 적합한 개체를 퍼트리는 과정이다. 적합도에 따라 개체를 선택할 확률을 정하는 방법에는 룰렛-휠 선택(Roulette-wheel selection), 순위 기반 선택(Ranking-based selection), 토너먼트 선택(Tournament selection) 등이 있다(진강규, 2000). MOVS는 이중에서 일반적으로 사용되는 룰렛-휠 선택을 사용한다. 크기가 s인 해집단에서 c번째 개체가 재생산 과정에서 선택될 확률은 다음과 같다.

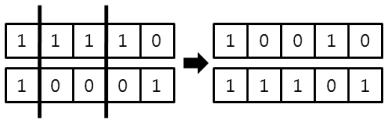


그림 2. 이점 교배의 예

$$P(select\ c) = \frac{Fitness(c)}{\sum_{i=1}^{s} Fitness(i)} \tag{16}$$

이때 $Fitness(c)$는 c번째 개체의 적합도이다.

(6) 교배와 돌연변이(S5, S9) : 교배 연산은 두 부모 유전자의 조합을 통해 새로운 자손을 생성하는 것이다. 교배 연산에는 일점교배(One-point crossover), 다점교배(Multi-point crossover), 균등교배(Uniform crossover) 등이 있다(진강규, 2000). MOVS는 선행 실험을 통해 다점교배 중 하나인 이점교배(Two-point crossover)를 이용하며 한 쌍의 부모 개체로부터 한 쌍의 자손 개체를 생성한다.

돌연변이 연산은 개체를 일정 확률로 임의 변환 시키는 연산이다. 돌연변이 연산을 통하여 개체군의 다양성을 유지할 수 있고 지역 최적해에서 벗어날 수 있다. 돌연변이 연산자는 단순돌연변이(Simple mutation)를 이용한다. 단순돌연변이는 돌연변이가 일어날 유전자를 선택하여 비트반전을 일으키는 것이다.

교배와 돌연변이는 확률적으로 발생한다. 알고리즘 내에서 교배 연산을 수행할 확률을 $P(c)$로, 돌연변이 연산을 수행할 확률을 $P(m)$으로 설정한다.

(7) 전체 기준 기반 평가 (S6) : 개체가 5가지 기준들을 얼마만큼 고루 충족시키는지를 평가한다. 평가 기준은 각 기준을 통해 나온 개체 순위의 합계를 사용한다. 개체 순위의 합계가 낮은 순으로 다시 순위를 결정해 식 (18)에 의해 적합도를 계산한다.

이 과정에서 순위가 1인 개체를 각 반복단계(Iteration)의 최적해로 판단한다. 알고리즘 종료 시 최종 반복단계의 최적해가 알고리즘이 찾은 최적해가 된다. 이 과정은 다중목적식에 적합한 개체들 중에서도 최종적으로 사용될 변수 조합을 정하는 역할을 한다. 또한 알고리즘의 종료조건을 판별하는데 사용된다.

(8) 종료조건 검사 및 결과 출력(S7) : 전체 기준 기반 평가 중 현재 반복단계에서 최선의 해와 전 단계에서 최

선의 해와 비교한다. 이때 최선의 해가 변하지 않은 반복 단계의 횟수(N)를 판단하여 지정한 횟수(E)만큼의 반복 단계에서 최선의 해가 변하지 않으면 알고리즘을 종료하고 결과를 출력한다.

4. 실험 및 분석

MOVS의 성능을 실험하기 위해서 두 가지의 실험을 수행하였다. 첫 번째는 다양한 조건을 가정하여 인위적으로 생성한 데이터를 이용한 LARS와의 비교 시뮬레이션이다. 이 실험을 통해 변수선택의 난이도가 다른 다양한 조건에서 유효변수 판별에 대한 두 알고리즘의 성능을 비교하였다. 단, 데이터의 수가 후보독립변수의 수보다 적은 경우는 LARS를 사용한 변수선택에 있어 명확한 사용 기준이 없기 때문에 비교실험 조건에서 제외되었다.

두 번째 실험은 포트폴리오 선택에 대한 과거의 주가 데이터를 이용한 시뮬레이션이다. 포트폴리오 선택에 관한 실험의 목적은 다음과 같다. 후보독립변수가 많은 실제 데이터에 MOVS를 적용시켜 응용문제 해결에 도움이 되는 것을 확인하고 알고리즘이 지닌 유연성을 입증하고 데이터의 수가 후보독립변수의 수보다 적은 경우에도 우수한 성능을 지님을 검증한다.

4.1. 생성한 데이터를 이용한 시뮬레이션

MOVS의 성능을 검증하기 위해 LARS와의 비교 실험을 수행하였다. 실제 데이터에서는 어떤 독립변수가 종속변수에 직접적으로 영향을 미치는지 알 수 없다. 이에 따라 본 논문은 종속변수에 영향을 미치는 독립변수를 정해준 상황에서 시뮬레이션을 수행한다. LARS에서 변수선택 기준은 C_p를 사용한다(Efron 등, 2004).

4.1.1. 데이터 생성 및 MOVS 매개변수 설정

실험 대상이 되는 모델은 선형회귀모델이다. 후보독립변수의 수는 100개(p = 100)이고 데이터의 수는 200개(n = 200)이다. 독립변수는 평균이 0, 표준편차가 1인 정규분포를 따르도록 생성한다. 오차항은 평균이 0, 표준편차가 $\lambda\sigma$인 정규분포를 따르도록 생성한다. σ는 오차항을 제외한 회귀식의 표준편차이고 λ는 상수이다. 유효한 독립변수들이 종속변수에 미치는 영향에 차등을 주기 위해 회귀계수 β는 1, 2, 3의 3가지 수준을 가지며, 각 회귀계수를 가지는 변수의 비율은 40%, 40%, 20%이다. 유효하지 않은 변수의 회귀계수는 0이다. 회귀계수가 0이 아닌 변수를 유효변수, 회귀계수가 0인 변수를 무효변수라 하겠다.

표 3. 실험에 사용한 MOVS 매개변수

매개변수	변수 설명	변수값
s	개체군 크기	20
$P(s)$	초기 해집단 설정에서 변수가 선택될 확률 (S1)	0.75
q	랭킹 가중치 (S3, S6)	0.1
$P(c)$	교차 확률 (S5, S9)	1
$P(m)$	돌연변이 확률 (S5, S9)	0.2
E	해가 변하지 않은 반복횟수에 따른 종료조건 (S7)	10

방해수준이 다른 여러 상황에 대하여 알고리즘을 평가하기 위해 조절한 요소는 두 가지이다. 하나는 후보독립변수 중 유효변수의 비율로 25%, 50%, 75%인 상황으로 나누었다. 다른 하나는 종속변수를 생성할 때 오차의 크기이다. λ값은 0.25, 0,45, 0.65의 3가지 수준을 가진다. 설정한 λ값에 따라 전체 후보변수를 이용하여 회귀모델을 구축하면 모델의 결정계수는 약 95%, 85%, 75%가 된다.

체계적인 매개변수 최적화는 하지 않았으나 간단한 선행실험을 통해 MOVS에 이용할 매개변수들을 표 3과 같이 동일하게 적용하였다.

4.1.2. 성능평가

표 4는 유효변수선택에 있어서 두 알고리즘의 성능을 비교한 것이다. FNR(Fault Non-discovery Rate)은 유효변수 중 선택된 변수 조합에서 빠져있는 유효변수의 비율로 유효변수를 제거하는 오류의 비율이다. FDR(Fault Discovery Rate)은 무효변수 중 선택된 변수 조합에 들어가 있는 무효변수의 비율로 무효변수를 선택하는 오류의 비율이다. TFR(Total False Rate)는 전체 변수 중 유효성 판별이 잘못된 변수의 비율로 전체적인 오류의 비율을 의미한다. 표에 나온 모든 값은 100번의 실험을 통해 산출된 평균값을 의미한다.

실험 결과를 보면 FNR은 (75%, 0.25), (75%,0.45)의 경우를 제외하고 MOVS가 LARS보다 작은 것을 확인할 수 있다. FDR과 TFR은 유효변수가 25%인 경우를 제외하고 MOVS가 LARS보다 작은 것을 알 수 있다.

변수선택에 있어 유효변수의 선택과 무효변수의 제거 중 하나에 초점을 맞추지 않는다면 우수한 변수선택의 기준은 전체적인 오류의 비율을 의미하는 TFR로 볼 수 있다. TFR만을 성능 평가의 기준으로 본다면 유효변수의 수보다 무효변수의 수가 적거나 같은 경우 MOVS의 성능이 LARS보다 우수하다. 그러나 유효변수보다 무효변

표 4. 변수 판별 성능 비교

		FNR		FDR		TFR	
유효변수	λ	LARS	MOVS	LARS	MOVS	LARS	MOVS
25%	0.25	0.001	0	0.118	0.148	0.089	0.111
	0.45	0.049	0.027	0.145	0.165	0.121	0.131
	0.65	0.15	0.095	0.142	0.178	0.144	0.157
50%	0.25	0.029	0.011	0.342	0.141	0.185	0.076
	0.45	0.15	0.117	0.253	0.171	0.202	0.144
	0.65	0.261	0.222	0.215	0.196	0.238	0.209
75%	0.25	0.036	0.059	0.585	0.173	0.174	0.088
	0.45	0.182	0.201	0.375	0.196	0.23	0.2
	0.65	0.328	0.303	0.283	0.211	0.317	0.28

표 5. 회귀모델의 통계량과 C_p 비교

		Adjusted R^2		p-value(F-test)		p-value(t-test)		SEE		C_p	
유효변수	λ	LARS	MOVS	LARS	MOVS	LARS	MOVS	LARS	MOVS	LARS	MOVS
25%	0.25	95.03	95.52	1.01E-89	1.07E-91	0.079	0.034	2.205	2.095	21.33	8.746
	0.45	85.64	87.12	1.23E-47	8.00E-57	0.1	0.042	3.962	3.754	21.46	8.297
	0.65	74.43	77.17	6.15E-36	3.04E-38	0.107	0.047	5.715	5.403	18.15	6.426
50%	0.25	94.63	95.23	1.59E-51	8.70E-78	0.119	0.018	3.26	3.073	65.09	40.11
	0.45	84.77	86.33	1.53E-35	2.41E-47	0.113	0.029	5.806	5.503	51.7	34.62
	0.65	73.28	76.04	1.50E-26	1.28E-28	0.133	0.04	8.312	7.872	41.33	27.47
75%	0.25	94.56	94.84	2.57E-48	8.81E-63	0.098	0.016	4.034	3.929	88.33	69.93
	0.45	84.2	85.41	1.16E-30	5.84E-38	0.119	0.027	7.269	6.989	72.69	56.31
	0.65	72.49	74.98	1.34E-23	7.99E-27	0.144	0.04	10.38	9.894	56.02	43.49

수가 많은 경우 LARS의 성능이 더 우수하다 할 수 있다.

하지만 이러한 현상은 FNR(유효변수의 선택)과 FDR 무효변수의 제거) 사이에 어느 정도 상충관계(trade-off)가 있기 때문에, 선택하는 변수의 수를 정하는 기준의 차이로 볼 수 있다. 즉 유효변수보다 무효변수가 많은 경우, LARS는 유효 변수의 수를 늘려가며 C_p가 가장 낮은 변수 조합을 선택하기 때문에 FNR은 MOVS보다 높지만 FDR을 줄여 전체적인 오류를 줄인 것으로 판단할 수 있다. 비슷한 예로 (75%, 0.25), (75%, 0.45)의 경우를 보면

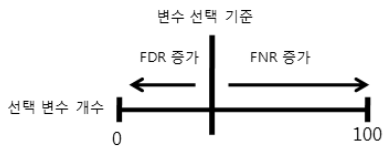


그림 3. FNR과 FDR의 상충관계

MOVS는 FNR이 LARS보다 높은 대신 FDR이 LARS보다 낮아서 TFR이 낮은 것을 알 수 있다.

FNR과 FDR의 상충관계를 배제하고 성능을 비교할 수 있는 예는 유효변수가 50%인 경우와 (75%, 0.65)의 경우에서 나타난다. 4개의 경우에 있어서 MOVS의 FNR과 FDR이 모두 LARS보다 낮은 것을 확인할 수 있다. 이 결과를 통해 MOVS가 LARS가 고려하지 못하는 변수조합을 고려함으로써 LARS보다 우수한 변수 조합을 찾아낸다는 것을 알 수 있다.

표 5는 회귀모델의 통계량과 C_p에 있어서 두 알고리즘의 성능을 비교한 것이다. 표에 나온 모든 값은 100번의 실험을 통해 산출된 평균값이다. 실험을 수행한 모든 경우를 검토한 결과 MOVS에 의한 변수 조합으로 구축한 회귀모델이 5가지 기준에 있어서 모두 유의성이 높은 것으로 나타났다. 즉, 일반적으로 사용하는 통계적 유의성에

대한 기준에 따르면, MOVS가 통계적 유의성이 더 높은 회귀모델을 구축한다고 볼 수 있다.

4.2. 주식 데이터를 이용한 포트폴리오 선택

포트폴리오 선택 문제에 있어서 주가(독립변수), 각 주식 종목에 투자할 자산의 비율(회귀계수)과 목표 수익률(종속변수)의 관계는 선형회귀모델로 표현될 수 있다. 즉, 주가와 목표 수익률이 있다면 선형회귀모델을 통해 투자할 자산의 비율을 정할 수 있다. 특히 통계적 유의성이 높은 회귀모델을 구축할 수 있다면 투자에 따른 위험 부담을 줄이는 포트폴리오 선택이 가능하다.

그러나 주식 종목의 수(p)가 예측에 사용할 수 있는 데이터의 수(n)보다 적기 때문에 주가 데이터에 회귀모델을 직접 적용시킬 수 없다. 또한 투자에 따른 위험 부담을 줄이기 위해서는 통계적 신뢰성이 높은 회귀모델이 필요하다. 따라서 회귀모델을 구축하기 위한 변수선택이 필요하고, MOVS를 이용하여 회귀모델을 구축하는데 사용할 주식 종목을 정한다.

4.2.1. MOVS의 변형

포트폴리오 선택에 MOVS를 사용하는데 있어서 고려해야 하는 특성은 4가지가 있다. 첫째, 데이터의 특성상 사용하는 변수의 수(p)가 데이터의 수(n)보다 많기 때문에 회귀모델을 구축할 수 없다. 둘째, 투자 비율이 음수가 될 수 없으므로 회귀계수가 음수인 변수는 제외시켜야 한다. 셋째, 데이터의 수가 부족하기 때문에 식 (16)에서 알 수 있듯이 *SSE*를 계산할 수 없고 C_p를 적합도 기준으로 이용할 수 없다. 마지막으로 포트폴리오 선택의 목표는 안정적인 수익의 증가이기 때문에 수익률 증가에 대한 영향이 확실하게 존재하는 종목만을 선택하는 것이 유리하다.

이와 같은 문제의 특성을 고려해 MOVS는 다음과 같은 변형을 거쳤다. 우선 표 2의 S2와 S6의 전 단계에서 개체 안에서 선택된 변수의 개수를 데이터의 개수 이하가 되도록 조정해준다. 이 과정에 있어서 각 변수가 제거될 확률은 동일하다. 그리고 S5와 S9의 전 단계에서 회귀계수가 음수인 변수를 제거한다. 끝으로 C_p를 사용하지 않고, 통계적 유의성이 높은 변수를 남기기 위해서 회귀모델의 평가 기준으로 F통계량의 p-value, 각 회귀계수에 대한 t통계량의 p-value의 평균값만을 사용한다.

4.2.2. 적용 방법과 결과

포트폴리오 구성에는 기준일로부터 30일의 주가 데이터를 사용한다. 30일당 목표 수익률은 30일간 종목별 주가의 평균 수익률과 수익률의 표준편차를 합한 값이다. 1일당 목표 수익률은 30일당 목표 수익률을 30으로 나눈 값이다. 특정 시점 t에서 i번째 주식의 주가를 p_{it}라고 정의한다면, 기준시점 T에서부터 시점 t까지, i번째 주식의 수익률 r_{it}는 다음과 같다.

$$r_{it} = \frac{p_{iT} - p_{it}}{p_{iT}} \tag{19}$$

기준일의 목표 수익률을 0으로 두고, 하루마다 1일당 목표 수익률을 더해서 30일간의 목표 수익률을 정한다. 정해준 목표 수익률을 종속변수로, 주가 데이터를 이용한 각 종목의 수익률을 독립변수로 설정하고 MOVS를 수행하여 투자할 종목을 선택한다. 선택된 투자 종목(선택 변수)을 이용하여 회귀모델을 구축하고 모델의 회귀계수를 이용하여 투자할 종목에 투자할 자산의 비율(w_i)을 정한다. w_i는 식 (20)와 같고, w_i의 합계는 1이다.

$$w_i = \frac{\beta_i}{\sum_{j=1}^{p} \beta_j} \tag{20}$$

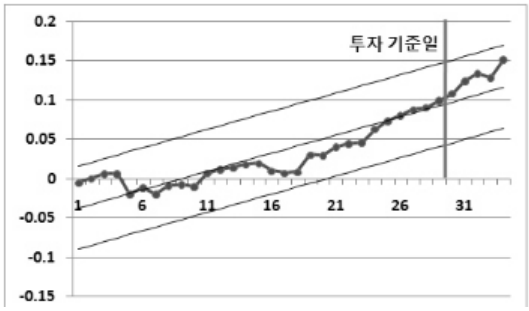
투자할 주식의 종목과 비율을 선택하면 투자 기준일로부터 이후 5일간 같은 방법으로 투자한다. 선택한 포트폴리오로 투자를 할 경우 기준시점 T에서부터 시점 t까지의 자산변화율 y_t는 다음과 같다.

$$y_t = \sum_{i=1}^{p} r_{it} w_i \tag{21}$$

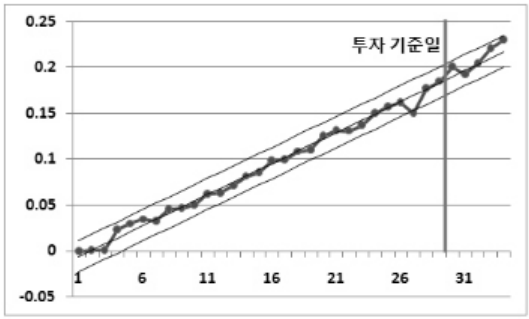
주가 데이터는 2006년 7월 20일부터 9월 28일까지 KRX (Korea Exchange, http://www.krx.co.kr)에서 거래된 실제 주식 정보 중, 주식 거래가 이루어지지 않은 날이 제외된 총 50일간의 종가(Closing price) 데이터를 사용하였다. 주식 종목은 결측데이터(Missing data)가 있는 종목을 제외한 565개를 사용하였다.

그림 4의 그래프는 회귀모델을 구축하는데 사용된 30일과 투자 기준일 이후 5일을 더한 총 35일간, MOVS를 통해 결정한 방법으로 투자할 경우의 자산 변화율을 나타내고 있다. (a), (b), (c), (d)의 경우에 있어서 선택한 투자 종목의 수는 각각 12, 18, 8, 10개이다. X축은 수익률을 계산하는 기준일로부터 지난 날짜이고, Y축은 자산변화율이다. 그래프에 나타난 적합선의 신뢰구간은 $\pm 3\sigma$이다.

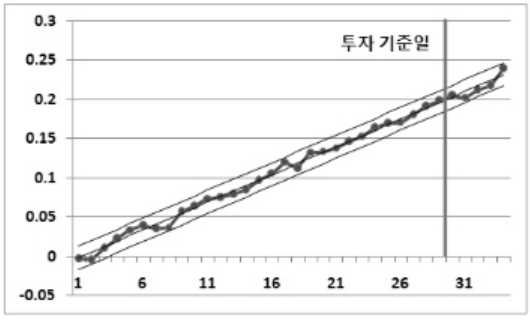
그림 4의 그래프에 따르면 자산이 적합선과 유사한 추



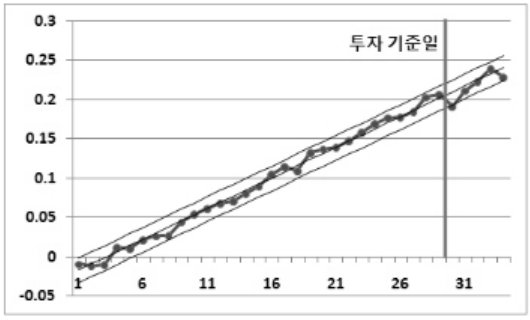
(a) 2006년 7월 20일~2006년 9월 7일



(b) 2006년 7월 27일~2006년 9월 14일



(c) 2006년 8월 3일~2006년 9월 21일



(d) 2006년 8월 10일~2006년 9월 28일

그림 4. 수익률 그래프

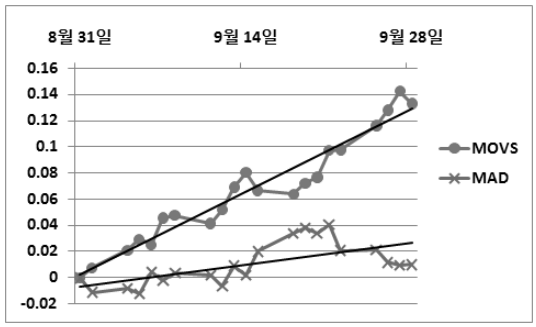


그림 5. MOVS와 MAD의 수익률 비교

세로 변화하는 것을 볼 수 있다. 그 결과로 MOVS를 응용한 포트폴리오 선택 방법은 자산이 꾸준히 증가하는 투자 종목과 투자율을 선택하는 것을 알 수 있다. 이러한 결과는 MOVS가 통계적 유의성이 높은 유효변수를 선택하고, 선택된 변수를 이용해 구축한 회귀모델의 회귀식과 회귀계수가 수익률을 예측하는 데 유효하다는 것을 의미한다.

MOVS를 응용한 포트폴리오 선택의 객관적인 성능을 검증하기 위해 일반적으로 사용되는 MAD 모형과 비교하였다. MAD 모형을 적용하는데 있어서 포트폴리오를 구성하기 위해 고려하는 기간의 길이는 MOVS와 같이 30일로 설정하였다. 그림 5는 앞서 언급한 데이터에 두 가지 포트폴리오 선택 방법을 적용시킨 결과로, 9월 1일부터 9월 28일까지, 20일간의 자산변화율의 그래프이다. 수익률의 기준일은 8월 31일이다.

그림 5에 나타난 그래프를 보면 MOVS를 응용한 투자가 MAD를 이용한 투자보다 자산증가율이 높은 것을 확인할 수 있다. 또한 적합선에 대한 표준편차는 MOVS가 0.0000947, MAD가 0.0001527로 MAD의 표준편차가 1.6배 이상 높게 나왔다. 그리고 적합선의 Adjusted R^2 값은 MOVS가 0.943, MAD가 0.429로 MOVS가 높게 나왔다. 이 결과는 자산의 변화 추세에 있어서 MOVS가 작은 변동폭을 가지고 있기 때문에 안정적인 투자를 한다는 것을 의미한다.

즉, MOVS를 응용한 포트폴리오 선택은 MAD와 비교하여 수익률이 높고, 위험부담이 작은 투자를 한다고 판단할 수 있다.

5. 결론 및 추후 연구

본 논문에서는 선형회귀모델을 위한 변수선택 알고리즘으로 다중목적 유전 알고리즘을 이용한 MOVS를 제안

하였다. MOVS는 회귀모델을 평가하는 다양한 기준을 만족시킬 수 있도록 설계되었다.

두 가지 실험을 통해 MOVS의 성능을 확인하였다. 다양한 조건에 대한 시뮬레이션을 수행한 결과, MOVS는 유효변수의 판별과 신뢰성 높은 모델링에 있어서 우수한 성능을 보였다. 그리고 주가데이터를 이용한 포트폴리오 선택 문제에 적용한 결과, MOVS의 응용을 통해 수익률이 높고, 위험부담이 작은 포트폴리오 선택이 가능한 것을 확인하였다. 두 가지 실험으로부터 MOVS는 변수 판별에 대하여 우수한 성능을 가지고, 현실의 문제를 해결하는 데 유효하며, 적용 상황에 따라 알고리즘을 유연하게 응용할 수 있음을 확인할 수 있었다.

MOVS를 발전시키기 위해서 MOVS에 사용하는 효과적인 매개변수와 유전 연산자에 대한 체계적인 연구가 필요하다. 또한 선형회귀모델 뿐만 아닌 다양한 모델들에 확대 적용시키는 연구를 통해 다양한 분야에 적용할 수 있을 것으로 기대된다.

참 고 문 헌

1. 김두섭, 강남중 (2008), *회귀분석 기초와 응용*, 나남, 파주.
2. 문병로 (2003), *유전 알고리즘*, 두양사, 서울.
3. 양경숙, 김미정 (2007), *R을 활용한 회귀분석*, 자유아카데미, 파주.
4. 진강규 (2000), *유전알고리즘과 그 응용*, 교우사, 서울.
5. Beasley J.E., Chu P.C. (1996), A genetic algorithm for the set covering problem. *European Journal of Operational Research*, Vol. 94, pp. 392-404.
6. Kalyanmoy Deb. (2001), *Multi-objective optimization using evolutionary algorithms, John Wiley & Sons, Ltd.*
7. Efron, B and T. Hastie (2004), “Least Angle Regression”, *The Annals of Statistics*, Vol. 32, No. 2, pp. 407-451.
8. Fonseca, C. M., Fleming, P. J. (1993). Genetic algorithms for multiobjective optimization : formulation, discussion and generalization. *Proceedings of the Fifth International Conference on Genetic Algorithms*. Morgan-Kauffman, pp. 416-423.
9. Furnival, G.M. and Wilson, R.W. (1974), “Regression by Leaps and Bounds”, *Technometrics*, Vol. 16, No. 4, pp. 499-511.
10. Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc.
11. T. Hesterberg, N. H. Choi, L. Meier, and C. Fraley (2008), Least angle and ℓ penalized regression : A review, *Statistics Surveys*, Vol. 2, pp. 61-93.
12. Isablle Guyon and Andre Elisseeff (2003), “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, Vol. 3, No. 10, pp. 1157-1182.
13. Konno, H., Yamazaki, H. (1991), “Optimization Model and Its Application to Tokyo Stock Market”, *Management Science*, Vol. 37, No. 5, pp. 519-531.
14. Luenberger, D. G. 1998. *Investment Science*. Oxford University Press, New York.
15. Mallows, C. L. (2000), “Some Comments on CP”, *Technometrics*, Vol. 42, No. 1, pp. 87-94.
16. Mitchell, T.M. (1997), Machine Learning, McGraw-Hill, Singapore.
17. T. Murata, H. Ishibuchi (1995), “MOGA : Multi-Objective Genetic Algorithms”, *Proc. of 2nd IEEE-ICEC Conferenc*, pp. 289-294.
18. C.R. Reeves (1993), *Modern Heuristic Techniques for Combinatorial Problems*, Blackwell Scientific.
19. J.D.Schaffer (1985), “Multiple objective optimization with vector evaluated genetic algorithms”, *the 1st International Conference on Genetic Algorithms*, pp. 93-100.

김 동 일 (dong22000@korea.ac.kr)

2008 고려대학교 공과대학 산업시스템정보공학과 학사
2008년~현재 고려대학교 정보경영공학전문대학원 석사과정

관심분야 : Advanced Process Control, Data Mining

박 정 술 (dumm97@korea.ac.kr)

2003 고려대학교 산업시스템정보공학과 학사
2006 고려대학교 산업시스템정보공학과 석사
2005~2006년 삼성경재연구소 6시그마실 Research Analyst
2006~2007년 고등기술연구원 로봇생산기술센터 연구원
2008년~현재 고려대학교 정보경영공학전문대학원 박사과정

관심분야 : Advanced Process Control, Data Mining

백 준 걸 (jungeol@korea.ac.kr)

1993 고려대학교 공과대학 산업공학과 학사
1995 고려대학교 공과대학 산업공학과 석사
2001 고려대학교 공과대학 산업공학과 박사
2001~2002 고려대학교 정보통신기술연구소 연구조교수
2002~2007 인덕대학 산업시스템경영학과 조교수
2007~2008 광운대학교 경영학부 조교수
2008년~현재 고려대학교 정보경영공학부 부교수

관심분야 : Advanced Process Control, Data Mining 응용, 지능형 이상진단

김 성 식 (sungskim@korea.ac.kr)

1972 고려대학교 공과대학 기계공학과 학사
1974 고려대학교 공과대학 산업공학과 석사
1979 미국 Southern Methodist University 산업공학 석사, 박사
1979~현재 고려대학교 정보경영공학부 교수

관심분야 : Advanced Process Control, Modeling, System Optimization